A flexible two-step randomised response model for estimating the proportions of individuals with sensitive attributes

Anne-Françoise Donneau, Murielle Mauer
Francisco Sartor and Adelin Albert

Department of Biostatistics, University of Liège, Liège
Scientific Institute of Public Health –Louis Pasteur, Brussels

Content

1. Problem definition
2. Classical RRM approach
3. New RRM approach
4. Properties
5. Application
6. Conclusion

Problem definition

Let p be the proportion of individuals in a given population presenting a certain characteristic of interest (e.g., smoking, heart defect, diabetes,…).

Problem: Estimating p from a sample of size n drawn from the population

Classical studies: Let r be the number of subjects with the characteristic in the sample, then an estimate of p is given by

$$\hat{p} = \frac{r}{n} \quad \text{and} \quad SE(\hat{p}) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$IC\ \ 95\%: \hat{p} - 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + 1.96\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Difficulties

When the attribute of interest is a “sensitive” one (e.g., illicit drugs consumption, psychiatric disorder,..), people tend to refuse to answer or intentionally mislead

↓

biased estimation of the true proportion p in classical method

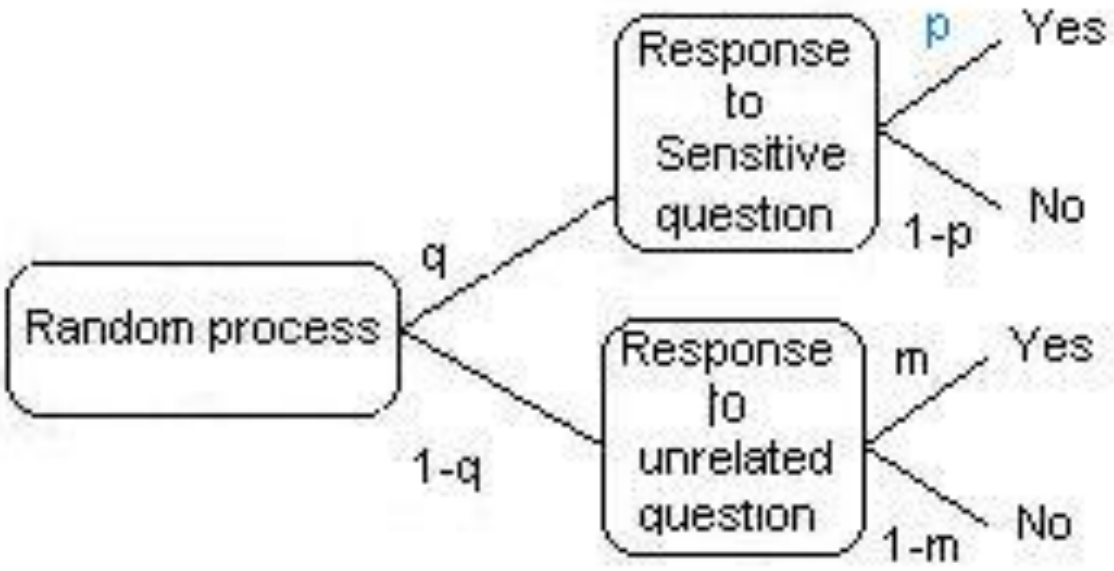
Solution to reduce or eliminate bias linked to the wish to keep private life secret →Random Response Model (RRM) (Warner, 1965)

RRM approach (I)

- Objective : reassuring the respondent that a possible affirmative answer to sensitive questions can't put him/her in a dangerous situation

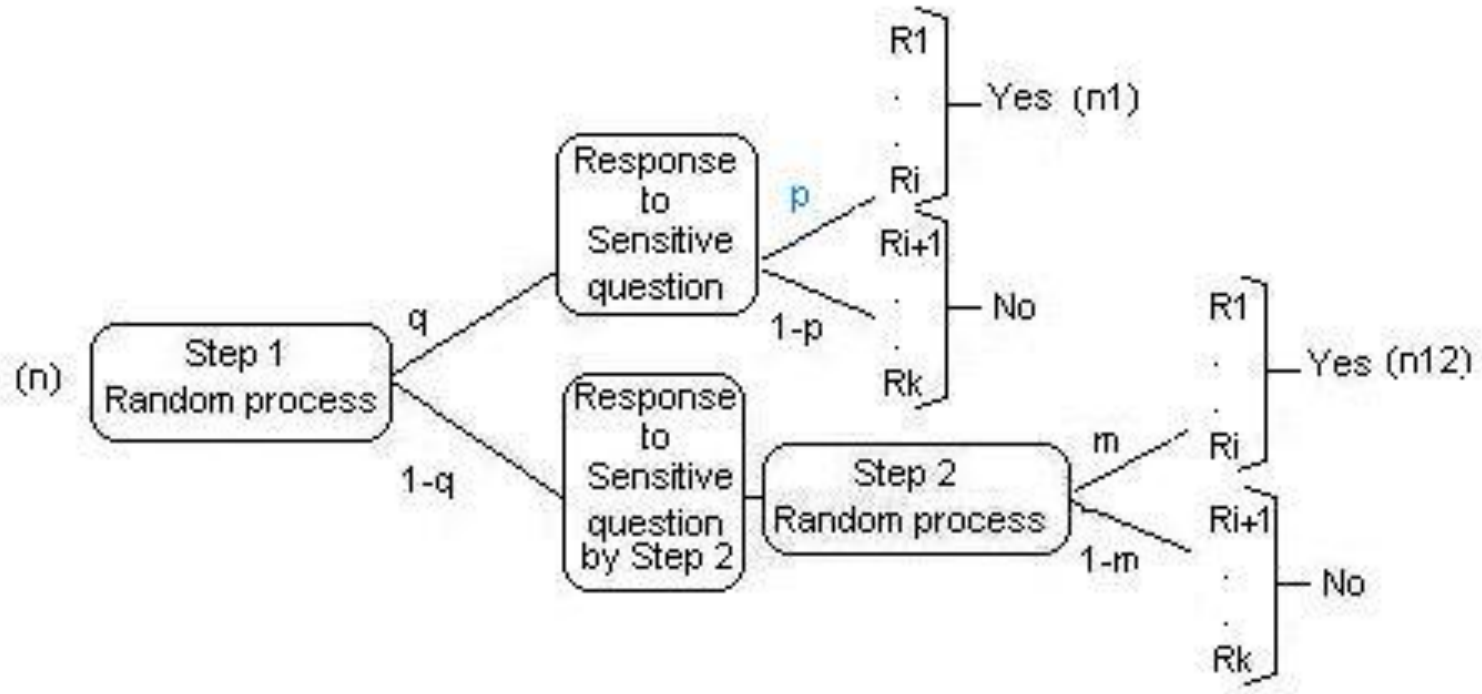
- Method: - Pair the sensitive question with an unrelated one.
  - Respondent select randomly one of these two questions.
  - Interviewer records only a "Yes or No" without knowing which question has been answered.

RRM approach (II)



New approach (I)

- No unrelated question but merely the sensitive question.
- Question with categorical answers which can be combined with some flexibility to have a binary outcome.



New approach (II)

Let f denote the probability of a positive answer regardless of the question,

$$f = qp + m(1-q)$$

Then, as the probabilities of the 2 random processes q and m are supposed to be known (and can be fixed a priori by the investigator), the probability of a positive answer to the sensitive question is

$$p = \frac{f - m(1-q)}{q}$$

Properties (I)

If n_1 and n_{12} are the observed numbers of subjects responding positively according to the outcome of the Step1 and Step1+ Step2, respectively

$$\hat{f} = \frac{n_1 + n_{12}}{n}$$

Consequently, $\hat{p} = \frac{\hat{f} - m(1-q)}{q}$ if the right hand-side is >0

and $\hat{p} = 0$ otherwise;

Properties (II)

Statement1: $\hat{p}$ is an unbiased and consistent estimator of p

- Unbiasedness $E(\hat{p}) = E\left[\dfrac{\hat{f} - m(1-q)}{q}\right]$

$$= \frac{E\left(\hat{f}\right) - m(1-q)}{q}$$

As $\hat{f}$ is an unbiased estimator of f

$$\rightarrow E(\hat{p}) = \frac{f - m(1-q)}{q} = p$$

- Consistency

By the law of large sample, $\hat{f} \rightarrow f$

Hence,

$$\hat{p} = \frac{\hat{f} - m(1-q)}{q} \rightarrow p$$

Properties (III)

Statment2 : $Var(\hat{p}) = \frac{1}{nq^2} f(1-f)$

$$Var(\hat{p}) = \left(\frac{\hat{f} - m(1-q)}{q} \right)$$

$$= \frac{1}{q^2} Var(\hat{f})$$

Since m and q are constant and as $Var(\hat{f}) = \frac{f(1-f)}{n}$

$$Var(\hat{p}) = \frac{1}{nq^2} f(1-f)$$

Properties (VI)

Thus for large n,

$$\frac{\hat{p}-p}{\sqrt{\frac{1}{nq^2}\hat{f}(1-\hat{f})}} \sim N(0,1)$$

Confidence interval

$$\left[\hat{p}-Q_Z(1-\frac{\alpha}{2})\sqrt{\frac{1}{nq^2}\hat{f}(1-\hat{f})};\ \hat{p}+Q_Z(1-\frac{\alpha}{2})\sqrt{\frac{1}{nq^2}\hat{f}(1-\hat{f})}\right]$$

One-sample power (I)

Determination of sample size n, so that $\left|\hat{p}-p\right| \le \Delta$ at the α% significance level

$$\Delta = Q_{z(1-\frac{\alpha}{2})}\sqrt{\frac{1}{nq^2}f(1-f)} \rightarrow n = \frac{Q_z^{2(1-\frac{\alpha}{2})}f(1-f)}{q^2\Delta}$$

As $f = qp + m(1-q)$,

$$n = \frac{Q_z^{2(1-\frac{\alpha}{2})}\left[qp+m(1-q)\right]\left[1-qp-m(1-q)\right]}{q^2\Delta}$$

One-sample problem (II)

Example based on the prevalence of the use of cannabis.

Hypothesis: $\Delta = 2\%$

$\alpha = 5\%$

$m = 5/6$ and $q = 0.5$

$p = 25\%$

Plugging these values in the previous formula, provides

$$\rightarrow \quad n \approx 191$$

Application : Question

During all life, how many times have you used illicit drug ?

1 = Never
2 = 1-2 times
3 = 3-5 times
4 = 6-9 times
5 = 10-19 times
6 = 20 times or more

better than

0=No
1=Yes

Application: Random process

Flip the coin (RP1), if the result is tail answer to question « a ». If the result is head, go to « b ».

a) During all life, how many times have you used illicit drug ?

1 = Never
2 = 1-2 times
3 = 3-5 times
4 = 6-9 times
5 = 10-19 times
6 = 20 times or more

Response ☐

b) Roll the dice once (RP2). What is the result ?

Application: Study material

Undergraduate students registered at the University of Liège during the academic year 2003-2004

Characteristics of the students (n=435)		
Age (year)		18.1 ± 0.78
Gender	Female (%)	288 (66.4)
	Male(%)	146 (33.6)

Application: Practical aspect

- During supervised practical works

  → more receptive and concentrated

- Explanation

- Closed box

Application: Results

Sensitive Question

"Yes" answer : 2-6 (1 or more)

"No" answer : 1 (Never)

$\longrightarrow$ m = 5/6

Sensitive question	Prevalence (± SE)	95% CI
Illicit drug	40.9 (± 4.7)	31.7 - 50.0

Application: Flexibility (I)

Sensitive Question

“Yes” answer : 4-6 (6 or more)

“No” answer : 1-3 (Never - 5 times)

$\rightarrow$ m = 1/2

Sensitive question	Prevalence (± SE)	95% CI
Illicit drug	26.7 (± 4.7)	17.6 – 35.9

Application: Flexibility (II)

Sensitive Question

“Yes” answer : 6 (20 or more)

“No” answer : 1-5 (Never - 19 times)

$\rightarrow$ m = 1/6

Sensitive question	Prevalence (± SE)	95% CI
Illicit drug	16.4 (± 3.6)	9.3 – 23.8

Application: Comparison anonymous questionnaire

A classical anonymous questionnaire was distributed to n = 462 undergraduate students also registered at the University of Liège during the academic year 2003-2004 .

Illicit drug	Prevalence (± SE)
Classical (n=462)	24.2 (± 2.0)
RRM (n=435)	16.4 (± 3.6)

Conclusion

- The problem of estimating the prevalence of sensitive attributes is quite common (public health, social life, economy,…)
- New approach of RRM by a two-step (RP1 – RP2) rather than by a one-step (RP1) random process response
- Flexibility with RP2 (question with multiple answers, modify)
- Can the two-step RRM be a substitute for the classical approach in general?
- Sensitivity analyses should be carried on .